HW1 Statistical Analysis

Problem description:

The study of **Social groups** and the **collective behaviors** of their members are hot topic not only in sociology, but also in computer science. In this homework, we attempt to perceive the semantics of social groups from collective social and behavioral information. Given the categories of the social groups and some features of collective social and behavioral information, our final goal is to test whether these features can distinguish the group categories and whether we could predict the relationship among different features.

Data

All the data is stored in one file, named **data.xlsx**.

The dataset describes the online group collected from QQ. We select 2040 online groups with corresponding information in 14 columns (denoted Col[1-14]):

Col[1-2]: online group name, group category. As you know, each QQ group has a group name to describe the semantics of the group. For both privacy and intuition, some characters of the names are masked by '*'. The descriptions of category are shown below in Table 1:

Table 1. Category description

Category	Theme	No.
1	Online Game	484
2	Stock Market	300
3	House & Living	196
4	School Alumni	425
5	Organization & Industry	635

Col[3-14]: 12 dimension features, they are group size, message number, friendship relational density, sex ration, average age, variance of age, geographical area, mobile conversation ratio, conversation number, no-response conversation ratio, night conversation ratio, images ratio

Experiment

1. Anova**(20 points)**
   a) Do the one-way ANOVA test for Col[7] with categories in Col[2]. Write down your conclusion, supporting statistics, and visualize your data which inspires the process.

2. Regression problems.
   a) **(20 points)** Only keep the data whose Col[11] >=20, let Col[12~14] be the labels, use average age (Col[7]) to predict Col[12~14] by linear regression. Show the linear function and calculate the prediction errors. What is the correlation between group average age and chatting behaviors? You need to plot the result of linear regression.
   b) **(20 points)** Let the number of conversations (Col[11]) as the weight of data. Let Col[12~14] be the labels, use 8 features(Col[3~10]) together to respectively predict Col[12], Col[13] and Col[14] by weighted multivariate linear regression. Show the linear function and calculate the prediction errors.

c) **(20 points)** Choose the data of categories 1 and 4, and then split the data into training set and validation set for binary classification (e.g., randomly sampling 20% data as the validation set and use the rest 80% as the training set). You can:

(1) classify training set by logistical regression using other features, and also test the model on the validation set,

(2) you can try to use different features and find which features can better predict the category,

(3) you can try multi-class classification on all categories.

Report your experimental settings and results.

3. **(20 points)** Redo one of the regression in question 2 a) by sampling 10% data (i.e. around 200 groups). Repeat 10 times and compute mean and standard deviation of the learned parameters(if you use 1-parameter regression, calculate the mean and variance of w_0 , w_1). Compare at least two sampling strategies. Which sampling method is more stable?How are the results compared to the results without sampling? Why?

Submission

You should submit one compressed file with:

1) One report, either in English or Chinese. It should be no more than 8 pages.
2) If you have codes (e.g. Python, Matlab, Java, etc.), put them into one file. Do not submit multiples files for codes. You don't need to submit codes if you use SPSS/excel/etc.
3) Do not submit other files, e.g., intermediate results. Figures should be included in the report rather than in separate files.
4) All files should be named with your ID, e.g., 2020001002.pdf, 2020001002.py, 2020001002.zip. Please check carefully (because we need to put your files into a system to check for plagiarism).
5) Failing to satisfy the above requirements will cost 5% of your grades.
6) Late submissions will cost 20% of your grades per week. The number of weeks is calculated by rounding up, e.g., if your submission is late by 1 day, it will be accounted for as $\lceil 1/7 \rceil$ = 1 week. **Start early!**
7) Each student should do his/her homework independently. Do not copy codes, reports, results, or any material from others, or share them with others, including online and publicly available sources, e.g. in Github. One exception is to use any build-in function in the software, e.g., in Excel, Python, Matlab. Note that both copying from others and intentionally providing materials for others to copy are considered plagiarism (so do not send your codes, results, reports to others) and will result in 0 grades and/or failing the class and/or other consequences instructed by the school honor code 《清华大学学生纪律处分管理规定实施细则》. It is a “red line” and we take it seriously